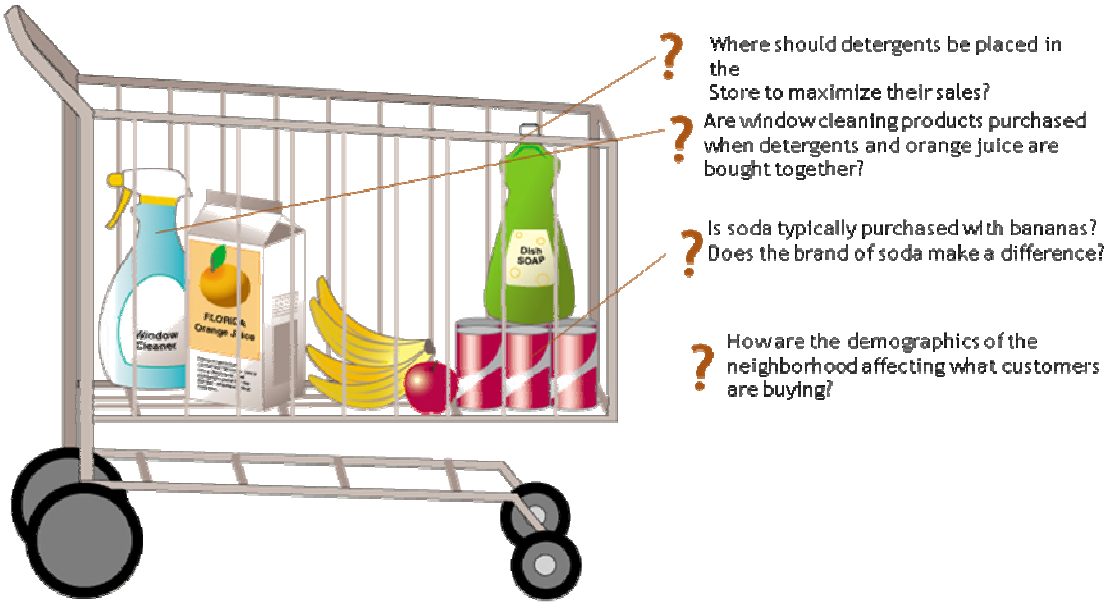
Chapter 9

Market Basket Analysis

9.1 關聯規則的定義和說明

—關聯規則(association rules):主要是從龐大資料中萃取出一系列變數或因子間的關係，即探索資料之間變數或項目間隱含的關係。關聯規則又稱為購物籃分析(Market Basket Analysis; MBA)。



分析這些交易資料如同在賣場觀察每一位顧客購物籃裡究竟買了什麼商品。每一個購物籃代表一為顧客在某個時間點的採購行為和一筆交易記錄。購物籃分析即是從這些看似相關卻又不盡相同的交易記錄中，找出潛在有用的關聯規則，以了解消費者購買行為的特定趨勢及慣行，進而應用於行銷、研發、供應鏈管理等相關決策上。例如：賣場商品的配置、銷售配貨、購物動線安排、產品定價及促銷與相關宣傳廣告等，不僅可提升顧客對於賣場的整體滿意度，也可提升賣場的銷售利潤。

—Customer Analysis

Market Basket Analysis uses the information about what a customer purchases to give us insight into who they are and why they make certain.

— Product Analysis

Market Basket Analysis given us insight into the merchandise by telling us which products tend to be purchased together and which are most amenable to purchase.

—There has been a considerable amount of research in the area of Market Basket Analysis (MBA). Its appeal comes from the clarity and utility of its results, which are expressed in the form *association rules*.

—Given data:

The collection data set usually is commercial daily transaction data, which each transaction contains a set of items.

— Find all rules $X \rightarrow Y$ that correlation the present of one set of items X with another set of items Y.

Example：

When a customer buys bread and butter, they buy mike 85% of the time.

While association rules are easy to understand, they are not always useful.Useful：

On Fridays convenience store customers often purchase diapers and beer together.

Trivial：Customers who purchase maintenance agreements are very likely to purchase large appliance.

Indexplicable: When a new super market opens, one of the most commonly sold item is light bulbs.

Example A grocery point of sale transactions

Customer	Product Items
1	Orange Juice, Soda
2	Milk, Orange Juice, Window Cleaner
3	Orange Juice, Detergent
4	Orange Juice, Detergent, Soda
5	Window Cleaner, Soda

Co-occurrence of products table

	Orange Juice	Window Cleaner	Milk	Soda	Detergent
Orange Juice	4	1	1	2	1
Window Cleaner	1	2	1	1	0
Milk	1	1	1	0	0
Soda	2	1	0	3	1
Detergent	2	0	0	1	2

— The co-occurrence table contains some example patterns

- Orange juice and soda are more likely to be purchased together than any other two products.
- Detergent is never purchased with window cleaner or mike.
- Mike is never purchased with soda or detergent.

— These simple observations are example of association and may suggest a formal rule like:

- If a customer purchased orange juice, then the customer also purchased soda.

9.2 關聯規則的衡量指標

關聯規則常利用支持度、信賴度和增益等三個衡量指標來分別表示規則的顯著性(significance)、正確性及價值，透過給定最小支持度(minimum support)和最小信賴度(minimum confidence)作為支持度與信賴度的門檻值(threshold)，再評估該條規則的資訊價值和增益。

若該條規則的支持度和信賴度大於或等於分析人員所訂定之門檻值，則表示該條規則有助於進行後續推論，若該條規則的增益值大於 1，即表示該條規則其發生的機率比原先的機率高，亦及該條規則有效。

關聯規則三項衡量指標的計算公式和意義說明如下：

1. 支持度(support)：衡量前提項目 (X) 與結果項目 (Y) 一起出現的機率，即 $P(X \cap Y)$，表示該條規則在全部交易紀錄中出現的比率。

   $Support\{X \Rightarrow Y\} = P(X \cap Y)$

   在例題 1 的交易紀錄中，若欲瞭解消費者購買 Orange juice 的同時也會購買 Detergent 的規則是否具有顯著性，可透過支持度衡量，即計算顧客同時會購買 Orange juice 和 Detergent 的機率，計算如下：

   $$Support(Orange\ \ juice \Rightarrow Detergent) = P(Orange\ \ juice \cap Detergent)$$
   $$= \frac{2}{5} = 40\%$$

2. 信賴度(Confidence)：衡量前提項目 (X) 發生的情況下，結果項目 (Y) 發生的條件機率，即 $P(Y \mid X) = \frac{P(X \cap Y)}{P(X)}$，表示對當前提項目 (X) 發生時，可推得結果項目 (Y) 的規則之正確性的信心程度。信賴度是衡量關聯規則是否具有可信度的指標；因此，信賴度須達到一定水準(通常為 0.5)，利用最小信賴度為門檻值去除正確機率較低的關聯規則。

   $$Confidence(X \Rightarrow Y) = P(Y \mid X) = \frac{P(X \cap Y)}{P(X)}$$

   在例題 1 的交易紀錄中，若欲瞭解消費者購買 Orange juice 的同時也會購買 Detergent 的規則之信心程度：

$$Confidence\{Orange\ juice \Rightarrow Detergent\} = \frac{P(Orange\ juice \cap Detergent)}{P(Orange\ juice)}$$
$$= \frac{2/5}{4/5} = 50\%$$

3. 增益(fit)：用於衡量比較信賴度與結果項目 (Y) 單獨發生時兩者機率的比值，即 $fit\,(X \Rightarrow Y) = \frac{P(Y \mid X)}{P(Y)}$。增益值的物理意義是比較關聯規則信賴度與原本結果項目 (Y) 發生之機率以衡量該規則之價值與相對效益，因此增益值至少要大於 1，表示該關聯規則的預測結果比原本表現好，亦即其信賴度大於原本結果項目 (Y) 發生之機率 。

$$Lift\{Orange\ \ juice \Rightarrow Detergent\} = \frac{P(Detergent \mid Orange\ juice)}{P(Detergent)}$$
$$= \frac{1/2}{2/5} = \frac{5}{4} = 1.25$$

進行關聯規則探勘時，通常會先設定探勘所得規則之支持度和信賴度的門檻值，以作為挑選關聯規則的準則。由此篩選出之規則必須滿足決策者訂定之最小支持度和最小信賴度。當滿足這兩個條件後，再判斷這些規則之增益值是否大於 1；大於 1 則保留，反之刪除。當三個指標皆成立時，即為所推導之關聯規則。

Example 2 某大賣場 5 位顧客的購買交易紀錄如下:

交易紀錄	商品(代碼)
101	牛奶(A)、麵包(B)、餅乾(C)、柳橙汁(D)
102	麵包(B)、餅乾(C)、汽水(E)、泡麵(F)
103	牛奶(A)、餅乾(C)、水果(G)
104	牛奶(A)、麵包(B)、柳橙汁(D)、泡麵(F)、水果(G)
105	餅乾(C)、汽水(E)、水果(G)

若分析人員設定支持度和信賴度門檻值分別為 0.2 和 0.5。

$$Support\{A, B \Rightarrow C\} = 0.2$$
$$Confidence\{A, B \Rightarrow C\} = 0.5$$
$$Lift\{A, B \Rightarrow C\} = \frac{0.5}{0.8} = 0.625$$

由於增益值為$0.625 < 1$，此規則「顧客於購買牛奶和麵包同時也會購買餅乾」，在經過最終衡量後，將不被列為顯著訊息，排除於有效的資訊集合中。

通常在欲探討之關聯規則中，商品項目越少時，消費該商品組合的顧客人次會相對提升，該規則的顯著性會越強烈。

關聯規則分析廣泛應用於零售業及大型賣場的資料探勘，藉由分析後所取得的資訊，得知顧客所偏好的商品和其它商品間的關聯，以制訂良好的市場行銷及配售計畫。

Example 3

Transaction ID#	Product Item
1	{1, 2, 3}
2	{1, 3}
3	{1, 4}
4	{2, 5, 6}

For minimum support 50%,

Frequent Item set	Support
{1}	75%
{2}	50%
{3}	50%

For the rule, if the customer purchased item1, then the customer purchased item 3.

$$Support\{1 \Rightarrow 3\} = P(1 \cap 3) = \frac{2}{4} = 50\%$$

$$Confidenc\{1 \Rightarrow 3\} = P(1 \mid 3) = \frac{2/4}{3/4} = 2/3 = 66.67\%$$

$$Life\{1 \Rightarrow 3\} = \frac{P(1 \mid 3)}{P(3)} = \frac{2/3}{2/4} = 1.33$$

9.3 關聯規則演算法

The Apriori Algorithm

— Apriori is an influential algorithm for mining frequent item sets for Boolean association rules.

— Apriori employs an iterative approach known as *level – wise* search, which $k-$item sets are used to explore $(k+1)$-item sets. First, the set of frequent 1-item sets is denoted L_1. L_1 is used to find L_2, and so on, until no more frequent k-item sets can be find. The finding of each L_k requires one full scan of the database.

Example

Transaction ID#	Item
1	{1,3,4}
2	{2,3,5}
3	{1,2,3,5}
4	{2,5}

Using the Apriori algorithm mining the association rule with minimum support is 75%.

Item sets	Support
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

Item sets	Support
{2}	3
{3}	3
{5}	3

L_1

Item sets	Support
{2,3}	2
{2,5}	3
{3,5}	2

Item sets	Support
{2,5}	3

L_2

Generate association rules:

1. If the customer purchased item 2, then the customer purchase item 5. (75%, 100%)
2. If the customer purchased item 5, then customer purchase item 2. (75%. 100%).

練習題

交易紀錄	商品(代碼)
201	牛奶(A)、麵包(B)、餅乾(C)、柳橙汁(D)
202	麵包(B)、餅乾(C)、汽水(E)、泡麵(F)
203	牛奶(A)、餅乾(C)、水果(G)
204	牛奶(A)、麵包(B)、柳橙汁(D)、泡麵(F)、水果(G)
205	餅乾(C)、汽水(E)、水果(G)